الإحصاء - المحاضرة السادسة – هندسة الروبوت

Statistics and probabilities- Lecture 6

Dr Fadi KHALIL

Doctor lecturer in statistics and programing

2021

في هذه المحاضرة سيتم التطرق لـ

بعض التوزيعات الإحتمالية المستمرة Some continuous probability distributions

- توزيع كاي مربع Chi squares distribution.

- توزيع فيشر Fisher distribution.

مفهوم الاحتمال الشرطي ونظرية Bayes:

مفاهيم إحصائية مهمة: مفهوم درجات الحرية *Degrees of freedom*:

يمكن تبسيط مفهوم درجات الحرية بأنّه مرتبط بحجم العينة و باستقلالية مفردات هذه العينة. أي أنّه يعبر عن عدد المفردات أو المشاهدات المستقلة التي تم استخدامها لحساب المؤشر الإحصائي (متوسط مشاهدات العينة مثلاً).

ماذا يعني استقلالية مفردات العينة؟

فرضاً إذا كان لدينا 5 مفردات في عينة ما وهي : 1 2 3 4 ؟ وكان متوسط هذه المفردات هو 3، من الواضح أنّ المفردة الخامسة المفقودة لا يمكن أن تبقى حرّة بل يمكن القول أن المفردة الخامسة هي حتماً 5 لكي يبقى المتوسط هو 3، بمعنى آخر أنّنا لا نملك الحرّية باختيار المفردة الخامسة. أي أنّ درجة الحرية صفر.

بينما إذا كان لدينا مفردتان مفقودتان 1 2 3 ؟ ؟ وكان المتوسط هو 5 هنا نلاحظ اننا نملك كامل الحرية لاختيار المفردة الرابعة المفقودة، ولكن بعد تحديدها سنكون مجبرين لقيمة واحدة فقط نعطيها للمفردة الخامسة لكي يبقى المتوسط هو 5، أي أنّ درجة الحرية هي درجة حرّة واحدة.

ونستمر إذا كان لدينا ثلاث قيم مفقود نلاحظ أنّه لدينا كامل الحرية لاختيار مفردتين مستقلتين، أي أنّه ليدنا 2 درجة حرية، وهكذا إلى أن نصل إلى حالة وجود 5 قيم مفقودة عندها يكون لدينا أربع قيم مفقودة. وبتعميم هذا المثال نلاحظ ان درجات الحرية يساوي عدد القيم ناقص واحد. هنا قمنا بطرح واحد لأنّه لدينا مؤشر إحصائي واحد نقوم بحسابه.

ولكن إذا كان لدينا أكثر من مؤشر إحصائي مثل حالة نموذج الانحدار، نقوم بطرح عدد معاملات الانحدار (أو بالبارمترات) في هذا النموذج. مثلاً إذا كان لدينا نموذج انحدار بثلاث معاملات تكون درجة الحرية هي عدد المشاهدات ناقص 3.

في نفس السياق عند حساب المتوسط الحسابي للعينة وفقاً للصيغة:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

فإنّ هناك n مشاهدة حرّة لحساب هذه الصيغة.

أمّا لحساب صيغة التباين للعينة:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

فإنّ عدد المشاهدات (أو القيم) الحرّة المستخدمة لحساب هذه الصيغة هو $n - 1$ وليس n ، إذ تمّ استخدام n القيمة لحساب المتوسط وبعد ذلك تمّ استخدام هذا المتوسط لحساب التباين. أي بمعنى آخر للحفاظ على قيمة المتوسط $\bar{x}$ هناك $n - 1$ قيمة حرّة بينما تبقى المشاهدة أو القيمة الأخيرة مقيّدة:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

إذاً مفهوم درجات الحرية يعبر عن كمية البيانات الحرّة المتبقية بعد حساب المؤشر الإحصائي، وذلك بعد حذف البيانات المقيّدة نتيجة حساب هذا المؤشر.

في نفس السياق لنفترض أنّه يتوفر نقطتين فقط لتقدير معادلة خط مستقيم، وبالتالي يجب استخدام النقطتين لتقدير القاطع *intercept* والميل *slope*، في هذا الحالة لا يوجد أي درجة حرية لأنّه تم استخدام النقطتين لتقدير القاطع والميل من دون أي بيانات اخرى تستخدم لدراسة جودة هذا النموذج، بمعنى آخر انّه عند تقدير هذا النموذج تم استخدام كل البيانات المتاحة. وهذا على خلاف مثلاً حالة توفر 1000 قيمة يتم استخدامها لتقدير القاطع والمستقيم، حينها يتوفر 998 قيمة حرّة تستخدم لدراسة مدى جودة معادلة خط المستقيم في التعبير عن هذه القيم.

لماذا مفهوم درجة الحرية مهم؟

إنّ مفهوم درجات الحرية هو مفهوم مهم جداً يؤثر على موثوقية التحليل الإحصائي وعلى قوة الاختبارات المستخدمة لدراسة دلالة التقديرات الإحصائية. كذلك يلعب مفهوم درجة الحريّة في تحديد عدد المقدرات التي يمكن تقديرها، حيث لا يمكن تقدير عدد كبيرة من المجاهيل باستخدام عدد قليل من درجات الحرية.

مما سبق يتضح أهمية وجود حجم كبير للعينة المدروسة، لأنّ كبر حجم العينة يساعد في زيادة درجات الحرية وبالتالي نملك الموثوقية لعميم النتائج المستخلصة من العينة على المجتمع الكامل، وبالمقابل لا يمكن تعميم تقدير مستمد من عدد قليل من درجات الحريّة على مجتمع كامل.

9-2-2-4- توزيع كاي مربع *Chi squares* :

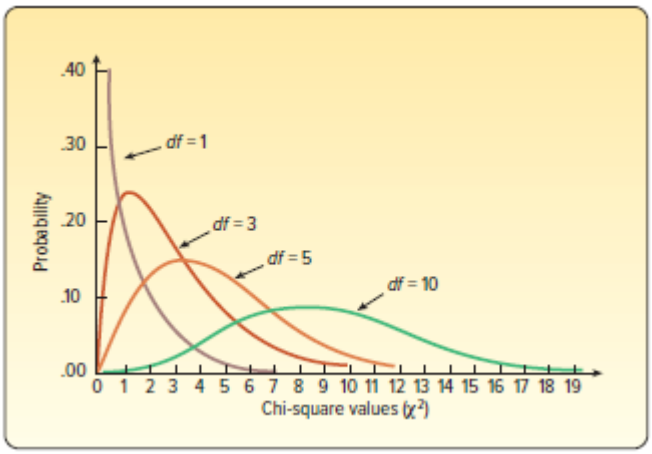
ليكن Z_1 متغير خاضع $N(0,1)$ فإنّ $X = Z_1^2$ يخضع توزيع كاي مربع بـ 1 درجة حرية.

ليكن Z_1 ، Z_2 ،..... Z_n ، n متغير عشوائي خاضع للتوزيع الطبيعي المعياري $N(0,1)$ ، بالتالي :

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

يخضع لتوزيع كاي مربع $X \sim \chi_n^2$ مع n درجة حرية degrees of freedom.

والذي يُعطى في الشكل الآتي:



من خواص توزيع كاي مربع:

1- إجمالي المساحة تحت المنحنى الإحتمالي تساوي 1.

2- يبدأ المنحنى من القيمة لصفر ويمتد باتجاه اللانهاية ويقترب من المحور الأفقي دون أن يلامسه.

3- منحنى توزيع كاي مربع ملتو نحو اليمين.

4- عندما تزداد درجات الحرية يصبح توزيع كاي مربع مشابهاً للتوزيع الطبيعي.

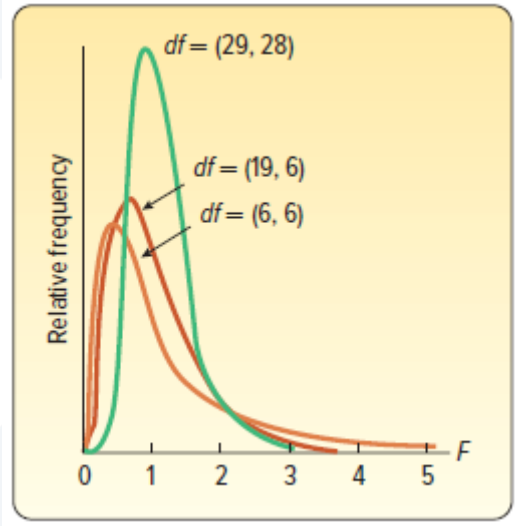
9-2-2-5- توزيع فيشر *Fisher distribution* :

إذا كان χ_n^2 و χ_m^2 متغيران خاضعان للتوزيع كاي مربع مع n و m دجة حرية على الترتيب. فإنّ المقدار :

$$\boldsymbol{F_{n/m}} = \frac{\chi_n^2/n}{\chi_m^2/m}$$

يخضع لتوزيع فيشر بدرجتي حرية، درجة حرية للبسط n و درجة حرية للمقام m.

والشكل الآتي يوضح المنحنى الإحتمالي لتوزيع فيشر.



من خواص توزيع فيشر :

1- يأخذ القيم الموجبة فقط من 0 إلى اللانهاية.

2- توزيع فيشر ملتو نحو اليمين *right skewed*،

3- كلّما ازدادت قيمة توزيع فيشر نحو قيم كبيرة فإنّ التوزيع الإحتمالي يقترب نحو المحور الأفقي دون أن يلامسه.

مفاهيم إحصائية مهمة: نظرية النهاية المركزية Central limit theorem:

تنص نظرية النهاية المركزية على أنّ مجموع عدد كبير من المتغيرات العشوائية المستقلة عن بعضها والخاضعة لنفس التوزيع الإحتمالي تقترب نحو التوزيع الطبيعي، وبشكل أكثر تحديداً:

بفرض X_1 ،X_2 ،..... X_n متغير عشوائي مستقلة عن بعضها وتخضع لنفس التوزيع الإحتمالي ولها متوسط μ وتباين σ^2 فإنّ توزيع المقدار، $X_1 + X_2 + \cdots X_n$ يخضع للتوزيع الطبيعي $N\sim(n\mu, n\sigma^2)$، بالتالي إن توزيع المقدار :

$$\frac{(X_1+X_2+\cdots X_n)-n\mu}{\sigma\sqrt{n}} \sim N(0,1)$$

وكذلك إنّ المقدار $\bar{X}=\frac{X_1+X_2+\cdots X_n}{n}$ يخضع للتوزيع الطبيعي $N\sim(\mu, \frac{\sigma^2}{n})$، أي:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

مثال :

يعتقد المهندسون المدنيون أن مقدار الوزن W (بوحدات 1000 رطل) يعبر عن مقدار الوزن الذي يمكن أن تتحمله مسافة معينة من الجسر دون حدوث أضرار هيكلية ، يتم توزيعه عادةً بمتوسط 400 وانحراف معياري 40. ومع افتراض أنّ الوزن (في وحدة 1000 رطل) من السيارة متغير عشوائي بمتوسط 3 وانحراف معياري 0.3 فكم عدد السيارات التي يجب أن تكون على طول الجسر حتى يتجاوز احتمال الضرر الهيكلي 0.1؟

بفرض أن وزن السيارة الواحدة هو متغير عشوائي X_i ، بالتالي إنّ الاحتمال :

$$P\left(\sum_{i=1}^{n} X_i > W\right)$$

$$P\left(\sum_{i=1}^{n} X_i - W > 0\right)$$

يعبر عن احتمال حدوث أضرار هيكلية في الجسر،

وأنّ $\sum_{i=1}^{n} X_i$ وفقاً لنظرية النهاية المركزية يخضع لتوزيع طبيعي بمتوسط $3n$ وتباين 0.09n ،

وكذلك المقدار $\sum_{i=1}^{n} X_i - W$ يخضع للتوزيع الطبيعي بمتوسط: $3n - 400$ ، وتباين $0.09n + 1600$

بالتالي :

$$P\left(\frac{\sum_{i=1}^{n} X_i - W - (3n-400)}{\sqrt{0.09n+1600}} > \frac{0-(3n-400)}{\sqrt{0.09n+1600}}\right)$$

$$P\left(\mathrm{z} > \frac{-3n+400}{\sqrt{0.09n+1600}}\right)$$

بحيث أنّ Z تخضع للتوزيع الطبيعي المعياري ، ووفقاً لمعطيات المسألة،

$$P\left(\mathrm{z} > \frac{-3n+400}{\sqrt{0.09n+1600}}\right) = 0.1$$

ومن جدول التوزيع الطبيعي المعياري : إن القيمة التي احتمالها 0.1 هي القيمة 1.28،

أي انّ $\frac{-3n+400}{\sqrt{0.09n+1600}} = 1.28$

أي انّ n تقريباً 117. بالتالي عند تجاوز عدد السيارات 117 هناك فرصة 0.1 لأن يحدث ضرر هيكلي في الجسر.

مثال :

قرر عالم فلك قياس المسافة بين مرصده و نجم بعيد. ولكن بسبب الأحوال الجوية ، فإن أي قياس لن يعطي المسافة الدقيقة d. نتيجة لذلك ، قرر الفلكي إجراء سلسلة من القياسات ثم استخدام قيمة المتوسط كتقدير للمسافة

الفعلية بين المرصد والنجم البعيد. إذا كان الفلكي يعتقد أن قيم القياسات المتتالية هي متغيرات عشوائية مستقلة بمتوسط d سنة ضوئية وانحراف معياري لسنتين ضوئيتين ، فكم عدد القياسات التي يجب أن يجعله متأكداً بنسبة 95٪ على الأقل من أن دقة تقديره في حدود ± 0.5 سنة ضوئية؟

الحل :

على فرض أن القياسات المأخوذ للمسافة هي متغيرات عشوائية مستقلة X_i ، بالتالي إنّ $\bar{X}$ يخضع للتوزيع الطبيعي بمتوسط d وتباين $\frac{4}{n}$

$$P(-0.5 \leq \bar{X} - d \leq +0.5)$$

$$P\left(\frac{-0.5}{2/\sqrt{n}} \leq \frac{\bar{X}-d}{2/\sqrt{n}} \leq \frac{+0.5}{2/\sqrt{n}}\right)$$

$$= 2P(z \leq +\sqrt{n}/4) - 1$$

ومن معطيات المسألة :

$$2P(z \leq +\sqrt{n}/4) - 1 \geq 0.95$$

$$P(z \leq +\sqrt{n}/4) \geq 0.975$$

في التوزيع الطبيعي المعياري $P(z \leq 1.96) = 0.975$

بالتالي $\frac{\sqrt{n}}{4} = 1.96$

أي على الأقل يجب ان يكون عدد القياسات 62 قياس للحصول على الدقة المطلوبة.

9-2-3- الاحتمال الشرطي Conditional probability:

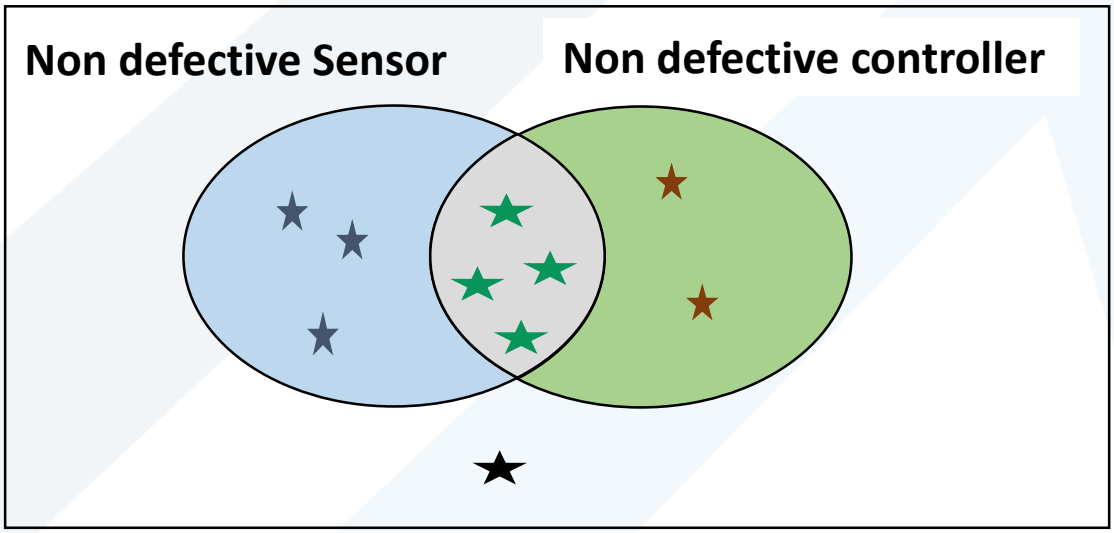
الاحتمال الشرطي يساعد أولاً على تضمين المعلومة المتوفرة عن نتيجة التجربة (حتى لو كانت ثانوية) في حساب الاحتمال المرغوب، وثانياً في إعادة حساب الاحتمال في حال توفر معلومة جديدة عن حدث آخر ثانوي مرافق للحدث الرئيسي.

ولتوضيح مفهوم الاحتمال الشرطي سيتم إيراد المثال الآتي:

مثال :

عند دراسة أداء 10 روبوتات ذكية تمّ فحص كل من مكون وحدة التحكم conroller و مكون المستشعرات Sensors وكانت النتيجة ممثلة في الجدول الآتي:

	Controller		
Sensors	Non defective	Defective	
Non defective	4	3	**7**
Defective	2	1	**3**
	6	**4**	**10**



لحساب احتمال وجود عيب في المستشعرات Sensors (DS)

$p(DS)=\frac{3}{10}=0.3$ ،

احتمال عدم وجود عيب في المستشعرات Sensors (NS)

$$p(NS)=\frac{7}{10}=0.7$$

أما احتمال وجود عيب في وحد التحكم Controller (DC)،

$$p(DC)=\frac{4}{10}=0.4$$

أما احتمال وجود عيب في وحد التحكم Controller (NC)،

$$p(DC)=\frac{6}{10}=0.6$$

يتم حساب هذه الاحتمالات **غير الشرطية** دون اللجوء إلى معلومات إضافية أو معلومات مسبقة تساعد في تحسين التنبؤ بوجود عيب سواء في المستشعرات أو وحدة التحكم.

مثلاً الجدول أو الرسم Venn يسمح بحساب احتمالات مشتركة joint probability:

احتمال عدم وجود عطل في وحدة التحكم والمستشعرات (اللون الاخضر) :

$$P(NS \cap NC) = \frac{4}{10}=0.4$$

احتمال عدم وجود عطل في وحدة التحكم مع وجود عطل في المستشعرات (اللون الاحمر) :

$$P(DS \cap NC) = \frac{2}{10} = 0.2$$

احتمال وجود عطل في وحدة التحكم مع عدم وجود عطل في المستشعرات (اللون الأزرق) :

$$P(NS \cap DC) = \frac{3}{10} = 0.3$$

احتمال أن وجود عطل في وحدة التحكم والمستشعرات (اللون الأسود) :

$$P(DS \cap DC) = \frac{1}{10}=0.1$$

ولكن عند أخذ معلومات إضافية بعين الاعتبار، مثلاً، احتمال أن يوجد عطل في المستشعرات مع العلم بعدم وجود عطل في وحدة التحكم، أي $P(DS/NC)$:

$$P(DS/NC) = \frac{2}{3} = 0.67$$

هنا يُلاحظ تحسن احتمال وجود عيب في المستشعرات مع أخذ بعين الاعتبار معلومة " **عدم وجود عطل في وحدة التحكم**".

كذلك عند حساب احتمال وجود عطل في المستشعرات ووحدة التحكم مع العلم المسبق بوجود عطل في لوحة التحكم:

$$P(DS \cap DC/DC) = \frac{1}{4} = 0.25$$

أيضاً يُلاحظ وجود تحسن في الاحتمال عند استخدام معلومة " **وجود عطل في لوحة التحكم**"

ولإضفاء مزيد من الوضوح على مفهوم الاحتمال الشرطي، سيتم أخذ الاحتمال وإجراء القليل من التعديلات، مثلاً قسمة البسط والمقام على 10 ليصبح:

$$P(DS \cap DC/DC) = \frac{\frac{1}{10}}{\frac{4}{10}} = 0.25$$

بالتالي يُلاحظ أن احتمال وجود عطل في المستشعرات ووحدة التحكم **بشرط بوجود عطل في لوحة التحكم** :

$$P(DS \cap DC/DC) = \frac{P(DS \cap DC)}{P(DC)} = 0.25$$

أي هو الاحتمال غير الشرطي لوجود عطل في المستشعرات ولوحة التحكم مقسوماً على الاحتمال غير الشرطي لوجود عطل في لوحة التحكم.

بنفس الأسلوب يمكن حساب الاحتمال الشرطي لوجود عطل في المستشعرات و وجود عطل في لوحة التحكم **بشرط وجود عطل في المستشعرات**

$$P(DS \cap DC/DS) = \frac{P(DS \cap DC)}{P(DS)} = 0.33$$

أي أنّ احتمال وجود عطل في المستشعرات مع وجود عطل في لوحة التحكم مع العلم المسبق بوجود عطل في المستشعرات هو 0.33.

مفهوم الاحتمال وفقاً لبايز Bayes's probability:

لشرح مفهوم الاحتمال البايزي بسهولة سيتم العودة للمثال السابق:

حيث تم إيجاد احتمال عطل في المستشعرات مع وجود عطل في لوحة التحكم بشرط (مع العلم المسبق) بوجود عطل في لوحة التحكم

$$P(DS \cap DC/DC) = \frac{P(DS \cap DC)}{P(DC)}$$

كذلك تمّ حساب الاحتمال الشرطي لوجود عطل في المستشعرات ووجود عطل في لوحة التحكم بشرط وجود عطل في المستشعرات

$$P(DS \cap DC/DS) = \frac{P(DS \cap DC)}{P(DS)}$$

السؤال المطروح هنا، هل يمكن إيجاد أحد الاحتمالين السابقين (الشرطيين) من دون معرفة الاحتمال غير الشرطي $P(DS \cap DC)$ ؟

بإصلاح الصيغتين السابقتين:

$$P(DS \cap DC) = P(DS \cap DC/DC) * P(DC)$$

$$P(DS \cap DC) = P(DS \cap DC/DS) * P(DS)$$

بما إنّ الطرف الأيسر هو نفسه:

$$P(DS \cap DC/DC) * P(DC) = P(DS \cap DC/DS) * P(DS)$$

بالتالي:

$$P(DS \cap DC/DC) = \frac{P(DS \cap DC/DS) * P(DS)}{P(DC)}$$

يلاحظ هنا إنّه أمكن حساب الاحتمال الشرطي من دون الحاجة إلى الاحتمال غير الشرطي $P(DS \cap DC)$.

وهذا ما يدعي بنظرية **بايز*Bayes's theorem***.

أي أنّ الاحتمال الشرطي لوجود عطل في المستشعرات ووحدة التحكم وبشرط (بالعلم المسبق) لوجود عطل في وحدة التحكم يمكن اشتقاقه من الاحتمال الشرطي $P(DS \cap DC/DS)$ والمرتكز على العلم بوجود عطل في المستشعرات.

ولتعميم العلاقة :

$$P(A \cap B/B) = \frac{P(A \cap B/A) * P(A)}{P(B)}$$

أي **الاحتمال الشرطي مع العلم بوجود معلومة عن الحدث يمكن اشتقاقه عن طريق معرفة معلومة أخرى عن هذا الحدث. بعبارة اخرى يمكن تضمين المعلومات التي نعرفها في حساب الاحتمال المرغوب.**

وتظهر أهمية هذه الصيغة عند عدم امتلاك المعلومات الكاملة عن الحدث المدروس، فمثلاً في المثال السابق يمكن حساب الاحتمال الشرطي لوجود عطل في المستشعرات ووحدة التحكم مع العلم المسبق بوجود عطل في لوحة التحكم عن طريق تقدير الاحتمال غير الشرطي لوجود عطل في لوحة التحكم، والاحتمال الغير الشرطي لوجود عطل المستشعرات، إضافة للاحتمال الشرطي لوجود عطل في المستشعرات ولوحة التحكم بشرط وجود عطل في المستشعرات.